ISSN 1561-5359. Штучний інтелект, 2019, № 1-2

УДК 004.89:004.93

М.С. Клименко
Інститут проблем штучного інтелекту МОН і НАН України, Україна
пр. Глушкова, 40, м. Київ, 03680

АЛГОРИТМ ПОШУКУ ЕЛЕМЕНТІВ МІЖФРАЗОВОГО ЗВ'ЯЗКУ У ПОСЛІДОВНОСТІ ВИСЛОВЛЮВАНЬ ТЕКСТУ

M.S. Klymenko
Institute of artificial intelligence problems MES and NAS of Ukraine, Ukraine
40, Hlushkov Ave., Kyiv, 03680

THE ALGORITHM FOR DEFINITION OF CONNECTIVE ELEMENTS BETWEEN PHRASES IN THE SEQUENCE OF TEXT STATEMENTS

У статті проаналізовано основні процедури пошуку елементів міжфразового зв'язку та розв'язання конфліктів посилань. На основі цього запропоновано узагальнений алгоритм, який комбінує переваги існуючих процедур пошуку елементів міжфразового зв'язку. Описано переваги обраних процедур та їх послідовності, наведено формальний опис вхідних даних та результатів алгоритму. Для оптимізації процедури сканування тексту алгоритм виконано у вигляді ітеративного зменшення кандидатів елементів зв'язку, що досягається за рахунок поступового підтвердження безконфліктних зв'язків.

Ключові слова: природна мова, семантичний аналіз, конфлікти посилань

In the article the basic procedures for finding of connective elements and resolving conflicts of references is analyzed. On the basis of this, a generalized algorithm is proposed that combines advantages of existing procedures for search for connective elements between phrases. The advantages of the selected procedures and their sequence are described, the formal description of input data and the results of the algorithm are presented. To optimize the procedure for scanning the text, the algorithm is performed as an iterative reduction of the candidates for communication elements. This is achieved through the confirmation of non-conflict links and the gradual resolution of conflicts.

Keywords: natural language, semantic analysis, conflict of references

Вступ

Семантичний аналіз тексту наразі знаходить застосування у широкому класі задач: тематична класифікація, анотування або стиснення тексту, переклад, пошук близьких за змістом документів, оцінка авторства, пошук знань та інші. Усі з наведених задач об'єднує спільна мета – необхідність виділення семантичних зв'язків для можливості обробки змісту тексту.

Роботи в області семантичного аналізу ведуться досить давно як вітчизняними, так і зарубіжними фахівцями. Однією з перших фундаментальних робіт можна вважати теорію породжувальної граматики [1], цілями якої було: побудова глибинної синтаксичної структури, запис значень кожного речення та виявлення семантичних аномалій. Розвитком в іншому напрямку стало створення предикатно-аргументних структур [2], в яких мовним конструкціям (аргументам) приписувалися ролі: агент, об'єкт, місце, адресат, інструмент, джерело. Серед розмаїття сучасних підходів до семантичного аналізу природномовних текстів практичного застосування знаходять методи на основі кластеризації та статистичного аналізу [3-7]. Дані методи використовуються для аналізу великих масивів даних (наприклад, веб-ресурсів або текстових бібліотек) оскільки мають відносно невелику обчислювальну складність. Однак, множина задач, які можуть вирішити дані методи, є вкрай обмеженою та потребує залучення експертів або додаткових засобів навчання систем [6-7]. Для задач, пов'язаних із трансформацією тексту (переклад, пошук знань), використовують модифікації нейронних мереж (наприклад, удосконалену довгу короткотривалу пам'ять [8]) або підходи, що засновані на онтологічних базах та засобах data mining [9, 10]. Недоліком зазначених вище засобів семантичного аналізу окрім нетривіальності та певної невизначеності, які неминуче виникають при використанні нейронних мереж, є також алгоритмічна складність невід'ємної факторизації матриць та тематична за-

лежність, які ускладнюють застосування даних засобів.

У даній роботі пропонується розробка універсального алгоритму пошуку елементів міжфразового зв'язку, який матиме змогу працювати за набором правил із довільним фрагментом природномовного тексту.

Мета дослідження

Метою даного дослідження є розробка детермінованого узагальненого алгоритму пошуку слів або словосполучень, які зв'язують за змістом вислови на природній мові.

Формування цілісного змісту тексту є розвитком роботи розпізнавання проявів емоцій людини за голосом [11]. Врахування семантичної ознаки тону висловлювання дозволить знизити помилку розпізнавання груп близьких за акустичними ознаками емоцій.

Постановка задачі

Із мети даної роботи випливають наступні задачі:

1. Визначити вхідні дані та вимоги до результату.
2. Проаналізувати основні процедури пошуку елементів міжфразового зв'язку та розв'язання конфліктів посилань.
3. Сформувати узагальнений алгоритм із обґрунтуванням обраних процедур та їх послідовності, описом переваг та недоліків даного підходу.

Опис дослідження

Для початку аналізу наявних елементів міжфразового зв'язку необхідно виконати побудову семантичної структури окремих висловлювань (простих речень або частин складних речень). Опис синтаксичної структури висловлювання можна виконувати шляхом виділення в ньому складових – груп слів, що функціонують як цілісні синтаксичні одиниці, або визначенням для кожного слова тих слів, які йому безпосередньо підпорядковані. У першому випадку використовується граматика складових і будується дерево складових, у другому випадку використовується граматика залежностей і будується відповідне дерево. Вхідними даними для алгоритму пошуку міжфразових зв'язків будемо вважати саме дерева залежності окремих висловлювань. Це зумовлено тим, що структура залежностей є більш інформативною для подальшого аналізу. Кореневим вузлом у дереві граматики залежностей виступає предикатор – логічний присудок висловлювання (яким зазвичай є дієслово із групи присудка).

Усі вирази, які входять до складу висловлювання, поділяють на дескриптивні та логічні терміни. *Дескриптивними термінами* називають слова або словосполучення, які позначають предмети, властивості, відношення чи дії, операції над предметами. *Логічними термінами* називають слова, які фіксують зв'язки, відношення, характеристики, що забезпечують інваріантність семіотичного інваріанту висловлювання за всіх можливих перетворень і будь-яких значень його дескриптивних термінів [12].

Формальний опис дерева залежностей буде мати наступний вигляд:

$$T = P(G_p, G_d, G_o),$$

де P – предикатор,
G_p – група підмета,
G_d – група додатка,
G_o – група обставини.

За сукупністю даних формул висловлювань створюється модель тексту, що представляє собою множину схем обчислення значень об'єктів, згаданих у висловах. Кожна така схема є своєрідним аналогом алгебраїчного дерева обчислень [6]. Отже, результатом виконання алгоритму пошуку елементів міжфразового зв'язку повинна бути описана вище модель тексту, яка є нелінійною комбінацією схем окремих висловлювань.

Пошук зв'язку між висловлюваннями можливо виконувати в прямому або зворотному порядку.

Прямий порядок передбачає виявлення референсних конструкцій, за якими аналізуються сусідні висловлювання та, за наявності сумісних із посиланнями груп підмету, додатку чи обставини, виконується об'єднання дерев семантичних залежностей. У якості референсних конструкцій виступає множина попередньо визначених термінів або послідовність лексичних одиниць, які сигналізують про наявність посилання у да-

ному висловлюванні на інший (частіше за все попередній) фрагмент тексту. Прикладами таких конструкцій є особові та вказівні займенники («той», «вони»), підсумовуючі («таким чином», «виходячи з цього») та порівняльні звороти з використанням слів («ще більше», «раніше»), прикметники у виняткових конструкціях («аналогічний») та інші. Зручність прямого порядку полягає в тому, що для референсних конструкцій можливо визначити групи термінів-адресатів, за якими виконуватиметься швидкий пошук у сусідніх висловлюваннях.

Зворотній порядок пошуку відштовхується від тези наявності зв'язку між усіма висловлюваннями. В такому разі постає необхідність попарного аналізу термінів та лексичних конструкцій на наявність зв'язку між ними. Перевагою зворотнього порядку є знаходження більшої кількості можливих зв'язків між висловлюваннями. Водночас це є й проблемою, оскільки нагальним питанням постає фільтрація зв'язків кандидатів та розв'язання конфліктів, які виникають внаслідок випадкових або лексично залежних співпадінь.

Опис алгоритму

У даній роботі запропоновано алгоритм пошуку елементів міжфразового зв’язку із поєднанням описаних вище підходів. Узагальнена схема алгоритму представлена на рисунку 1.

Слід зазначити, виникнення конфліктів між кандидатами елементів семантичних зв'язків між висловлюваннями виникає не тільки при зворотньому, а також можливо і при прямому порядку пошуку. Оскільки розв'язання конфліктів не є тривіальним завданням, необхідним буде мінімізація їх виникнення для прискорення роботи алгоритму.

Для оптимізації процедури сканування тексту алгоритм виконано у вигляді ітеративного зменшення кандидатів елементів зв'язку. Це досягається за рахунок підтвердження безконфліктних зв'язків та поступового розв'язання конфліктів.

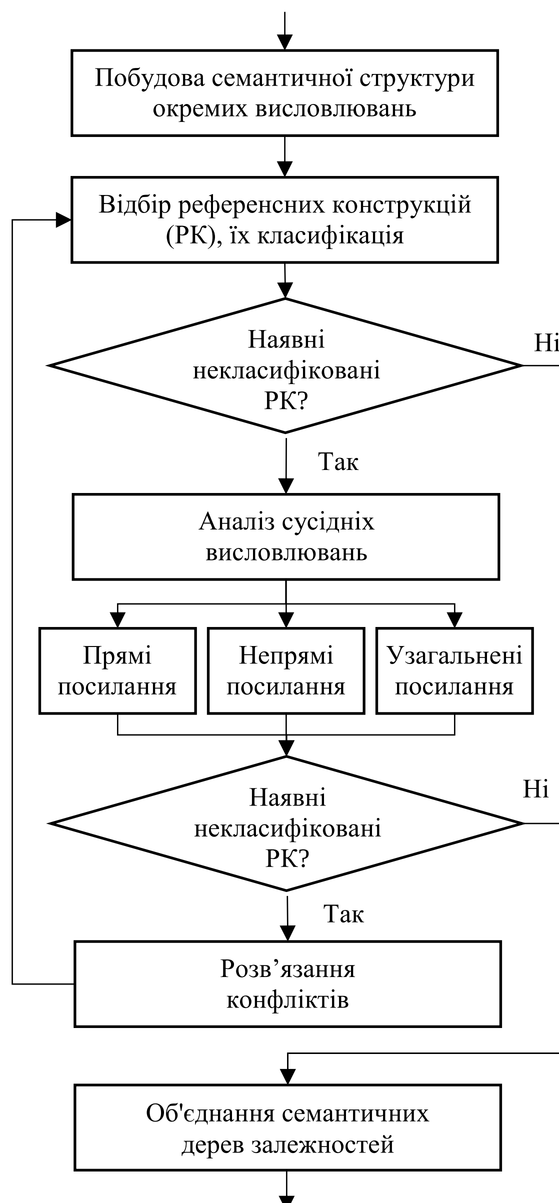


Рис. 1. Узагальнена схема алгоритму пошуку елементів міжфразового зв’язку.

На наступному етапі виконується аналіз сусідніх словосполучень, виходячи із типу класифікованих референсних конструкцій.

Для прямих посилань знаходяться узгоджені із ними слова та словосполучення. У разі знаходження декількох адресатів серед них обирається головніший за семантичним деревом залежностей, а якщо вузли-адресати однакової глибини в деревах, то конфлікт розв'язується аналогічно непрямим посиланням на наступній ітерації.

Для узагальнених посилань знаходяться дії або інші логічні предикати у сусідніх висловлюваннях. За їх відсутності або за

наявності конфліктів розв'язання виконується в рамках непрямих посилань наступним чином.

Визначення адресатів непрямих посилань виконується за допомогою наявної семантичної бази. Оскільки в даному випадку між претендентами на зв'язок між висловлюваннями відсутня узгодженість та близькість за синтаксичними деревами залежності, відбір адресатів відбувається за рахунок пошуку посилань на дані терміни у семантичній базі. Важливим є абстрагування текстових термінів від опису семантичних висловлювань усередині бази для вирішення проблеми синонімії та подальшого розширення семантичної бази іншомовною термінологією. Далі пошук звужується після проведення фільтрації за спільними атрибутами семантичної бази. Визначення семантичного зв'язку T між терміном K_T та референсним посиланням $K_x \in O$ має вигляд:

$$T = \min_{1 \le i \le n\ i <> T} D(K_T, K_x),$$

де n – кількість конфліктних конструкцій K_x.

Відстань D обчислюється як довжина між відповідними вузлами мережі семантичної бази.

Процес розв'язання конфліктів за рахунок фільтрації атрибутів у семантичній базі є найбільш ресурсомістким у даному алгоритмі, фактично виконується зворотній пошук зв'язку між висловлюваннями. Тому пріоритет надається знаходженню тривіальних зв'язків за словниками або узгодженими конструкціями і тільки в разі виникнення складних конфліктів залучається семантична база.

Висновки

У статті запропоновано удосконалений алгоритм пошуку елементів міжфразового зв'язку за рахунок комбінації підходів аналізу висловлювань та оптимізації їх послідовності виконання. Продовженням даної роботи може стати програмна реалізація та числове дослідження запропонованого алгоритму, порівняння ефективності результатів його роботи із результатами провідних алгоритмів.

Застосування семантичного аналізу послідовності висловлювань може підвищити ефективність прикладних задач розпізнавання образів. Зокрема, використання алгоритму у системі класифікації емоційних проявів за голосом гіпотетично може дозволити розпізнавати додаткові тональності емоцій, які наразі не можуть бути визначені виключно акустичними та просодичними характеристиками [11].

Література

1. Chomsky, N. (1957). *Syntactic Structures*. London: Mouton.
2. Fillmore, Ch. (1968). *The Case for Case. Universals in Linguistic* Theory. New York.
3. López-Quintero, J.F., Cueva Lovelle, J.M., González Crespo, R. (2018, March 22). A personal knowledge management metamodel based on semantic analysis and social information. Soft Computing (6, pp. 1845-1854). URL: https://doi.org/10.1007/s00500-016-2437-y
4. Yan W., Liu H., Liu Yu., Wang J., Zanni-Merk C., Cavallucci D., Yan X., Zhang L. (2018). Latent semantic extraction and analysis for TRIZ-based inventive design. *European Journal of Industrial Engineering* (12:5, pp. 661-681). URL: https://doi.org/10.1504/EJIE.2018.094593
5. Braun, V., Clarke, V., Hayfield, N., Terry, G. (2019). Thematic Analysis. In: Liamputtong P. (eds) Handbook of Research Methods in Health Social Sciences. Springer. Singapore. URL: https://doi.org/10.1007/978-981-10-5251-4_103
6. Ali, I., Melton, A. (2018). Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory. IEEE 12th International Conference on Semantic Computing (pp. 243-247). doi: 10.1109/ICSC.2018.00042
7. Garten J., Hoover J., Johnson K.M., Boghrati R., Iskiwitch C., Dehghani M. (2018, February). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior Research Methods (50, 1, pp. 344–361). URL: https://doi.org/10.3758/s13428-017-0875-9
8. Dangovski, R., Jing, L., Nakov, P., Tatalović, M., Soljačić, M. (2019). Rotational unit of memory: a novel representation unit for RNNs with scalable applications. Transaction of the Association of Computational Linguistics (7, pp. 121-138).
9. Gu, T., Wang, X.H., Pung, H.K., Zhang, D.Q. (2004). An ontology-based context model in intelligent environments. *Proceedings of communication networks and distributed systems modeling and simulation conference*.
10. Guo, K., Liang, Zh., Tang, Ya., Chi, T. (2018). SOR: An optimized semantic ontology retrieval algorithm

for heterogeneous multimedia big data. *Journal of Computational Science* (28, pp. 455-465). URL: https://doi.org/10.1016/j.jocs.2017.02.005
11. Клименко, М.С. (2018). Удосконалений метод розпізнавання емоційного стану диктора із семантичним аналізом змісту. *Штучний інтелект*, 1, 22-27.
12. Конверський, А.Є. (2008). *Логіка (традиційна та сучасна)*. Київ: Центр учбової літератури.
13. Hofmann, T. (2017, August). Probabilistic Latent Semantic Indexing. SIGIR Forum 51, 2, 211-218. DOI: https://doi.org/10.1145/3130348.3130370
14. Ben-Or, M. (1983). Lower Bounds For Algebraic Computation Trees. Proc. 15th ACM Annu. Symp. Theory Comput, 80-86.
15. Perper, E.M., Gasanov, È.È., Kudryavtsev, V.B. (2018). On the semantic analysis of juridical documents. Intelligent systems. Theory and applications, 22:3, 45–88.
16. Xu, Z. et al. (2017). Hierarchy-Cutting Model Based Association Semantic for Analyzing Domain Topic on the Web. IEEE Transactions on Industrial Informatics (13:4, pp. 1941-1950). doi: 10.1109/TII.2017.2647986

References

1. Chomsky, N. (1957). *Syntactic Structures*. London: Mouton.
2. Fillmore, Ch. (1968). *The Case for Case. Universals in Linguistic Theory*. New York.
3. López-Quintero, J.F., Cueva Lovelle, J.M., González Crespo, R. (2018, March 22). A personal knowledge management metamodel based on semantic analysis and social information. Soft Computing (6, pp. 1845-1854). URL: https://doi.org/10.1007/s00500-016-2437-y
4. Yan W., Liu H., Liu Yu., Wang J., Zanni-Merk C., Cavallucci D., Yan X., Zhang L. (2018). Latent semantic extraction and analysis for TRIZ-based inventive design. European Journal of Industrial Engineering (12:5, pp. 661-681). URL: https://doi.org/10.1504/EJIE.2018.094593
5. Braun, V., Clarke, V., Hayfield, N., Terry, G. (2019). Thematic Analysis. In: Liamputtong P. (eds) Handbook of Research Methods in Health Social Sciences. Springer. Singapore. URL: https://doi.org/10.1007/978-981-10-5251-4_103
6. Ali, I., Melton, A. (2018). Semantic-Based Text Document Clustering Using Cognitive Semantic Learning and Graph Theory. IEEE 12th International Conference on Semantic Computing (pp. 243-247). doi: 10.1109/ICSC.2018.00042
7. Garten J., Hoover J., Johnson K.M., Boghrati R., Iskiwitch C., Dehghani M. (2018, February). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior Research Methods (50, 1, pp. 344–361). URL: https://doi.org/10.3758/s13428-017-0875-9
8. Dangovski, R., Jing, L., Nakov, P., Tatalović, M., Soljačić, M. (2019). Rotational unit of memory: a novel representation unit for RNNs with scalable applications. Transaction of the Association of Computational Linguistics (7, pp. 121-138).
9. Gu, T., Wang, X.H., Pung, H.K., Zhang, D.Q. (2004). An ontology-based context model in intelligent environments. *Proceedings of communication networks and distributed systems modeling and simulation conference*.
10. Guo, K., Liang, Zh., Tang, Ya., Chi, T. (2018). SOR: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data. Journal of Computational Science (28, pp. 455-465). URL: https://doi.org/10.1016/j.jocs.2017.02.005
11. Klymenko, M.S. (2018). Udoskonalenyj metod rozpiznavannya emocijnogo stanu dyktora iz semantychnym analizom zmistu. *Shtuchnyj intelekt*, 1, 22-27.
12. Konverskyj, A.Ye. (2008). Logika (tradycijna ta suchasna). Kyiv: Centr uchbovoyi literatury.
13. Hofmann, T. (2017, August). Probabilistic Latent Semantic Indexing. SIGIR Forum 51, 2, 211-218. DOI: https://doi.org/10.1145/3130348.3130370
14. Ben-Or, M. (1983). Lower Bounds For Algebraic Computation Trees. Proc. 15th ACM Annu. Symp. Theory Comput, 80-86.
15. Perper, E.M., Gasanov, È.È., Kudryavtsev, V.B. (2018). On the semantic analysis of juridical documents. Intelligent systems. Theory and applications, 22:3, 45–88.
16. Xu, Z. et al. (2017). Hierarchy-Cutting Model Based Association Semantic for Analyzing Domain Topic on the Web. IEEE Transactions on Industrial Informatics (13:4, pp. 1941-1950). doi: 10.1109/TII.2017.2647986

RESUME

M.S. Klymenko

The algorithm for definition of connective elements between phrases in the sequence of text statements

The object of this study is to develop a determined generalized algorithm for finding words or phrases that semantically connect the content of expressions in the natural language.

In the article the basic procedures for finding of connective elements and resolving conflicts of references is analyzed. Semantic text analysis is currently used in a wide variety of tasks: thematic classification, annotation or compression of the text, translation, search for related documents, authorship evaluation, knowledge search, and others. Works in the

field of semantic analysis are conducted for a long time by domestic and foreign specialists.

Among the variety of modern approaches to semantic analysis of natural language texts for practical application are methods based on clustering and statistical analysis. The range of tasks that these methods can solve are extremely limited and require the involvement of experts or additional training systems. For tasks related to the transformation of text, modifications of neural networks or approaches based on ontological databases and data mining are used. The disadvantage of these methods of semantic analysis is non-triviality and uncertainty in decision making that are inevitably arises in the application of neural networks, as well as the algorithmic complexity of negative matrices factorization and thematic dependency that complicates the use of these methods.

In this article the algorithm of search of connective elements between phrases with a combination of the described direct and reverse approaches is offered. The generalized scheme of the proposed algorithm is described.

Since the resolution of conflicts is not a trivial task, it will be necessary to minimize their occurrence in order to accelerate the algorithm's operation.

The future work in this direction may consists of program implementation and numerical research of the proposed algorithm, comparing the effectiveness of its results with the results of described modern algorithms.

Надійшла до редакції 13.05.2019